MANGALORE UNIVERSITY

Final Synopsis

of the PhD Thesis to be submitted for the award of degree of
Doctor of Philosophy in Computer Science

TITLE OF THE WORK

RGB-D Video Analytics for Human Action Recognition

By
Rathnakara Shetty P.,
Research Scholar

Under the Guidance of
Dr. B.H. Shekar,
Professor

Department of Computer Science,
Mangalore University,
Mangalagangotri-574 199

December 2021

Contents

1 Introduction **1**
- 1.1 Human Action Recognition . . . 1
- 1.2 Human Action Datasets . . . 2
- 1.3 Review of Literature . . . 3
- 1.4 Evaluation Process . . . 4
- 1.5 Motivation and Objectives . . . 6

2 UDTCWT based Descriptor **7**
- 2.1 Proposed Methodology . . . 7
- 2.2 Experimental Analysis . . . 10

3 Complex Gradient based Descriptor **11**
- 3.1 Proposed Methodology . . . 11
- 3.2 Experimental Results . . . 13

4 Stacked Filter Bank based Descriptor **15**
- 4.1 Proposed Methodology . . . 15

5 Neural Network based Model **19**
- 5.1 Proposed Methodology . . . 19
- 5.2 Experimental Analysis . . . 20

6 Comparative Analysis **21**
- 6.1 Performance Analysis . . . 21

7 Conclusion **23**
- 7.1 Contributions in our work . . . 25
- 7.2 Publications: . . . 26

RGB-D Video Analytics for Human Action Recognition

1 Introduction

Video based human action recognition is the process of labeling sequence of images or video frame sequences with action labels. Human activity recognition (HAR) aims to recognize activities in a given video sequence based on a sequence of observations of the movements or actions of subjects and the environmental conditions. Video based action recognition research has become the basis for many computer vision based applications such as gait recognition, video surveillance, healthcare, sign language interpretation, scene understanding, crime scene analysis, sports and athletic analysis and also human-computer interaction (HCI).

Automatic human action understanding is essential to build a human computer interfaces which can intelligently serve people. The process of identifying actions and gestures by humans in video sequences is called Action Recognition. This interpretation of human or robot induced actions and activities have gained considerable interest in the computer vision, entertainment industry, robotics and AI communities.

1.1 Human Action Recognition

Human activities can be categorized into three levels, with four different types where at the bottom level we find the atomic elements called action primitives. After the action primitive level, we find action/activity constituted with a set of action primitives. At the top level we find the complex interactions, which refer to the human activities that involve more than two persons and objects. Action primitives are the atomic actions at the limb level, such as *stretching the left arm*, and *raising the right leg*. Most of the symbolic gestures are atomic actions. Atomic actions are usually limited to a particular part of human body, such as arms, upper body part or hands. Normally, activity refers to the movements which involves several action primitives where the whole body is involved in the movement and moreover, it involves a temporal sequential order and only one person is involved in it. A trajectory of various representational methods used in HAR related research is shown in Figure 1.

Challenges in Action Recognition: Given an activity may vary in terms of speed, angle and style from subject to subject, which leads to the variations within the class, namely intraclass variations. On the other hand, sometimes two different activities possess similar movements and shapes. This is called as inter-class similarity. Most of the real-world videos have complex dynamic backgrounds. And, also most of such videos are captured using a dynamic video recording devices. Such broadcasts and videos are normally recorded in varying backgrounds. Earlier research was limited to low-level and simple human activities such as running, jumping and waving hands which were having

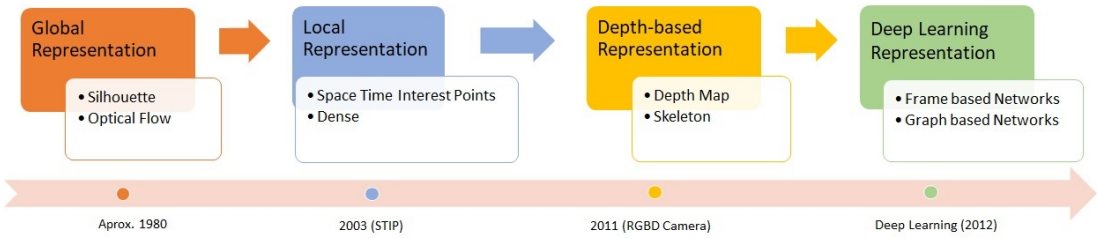


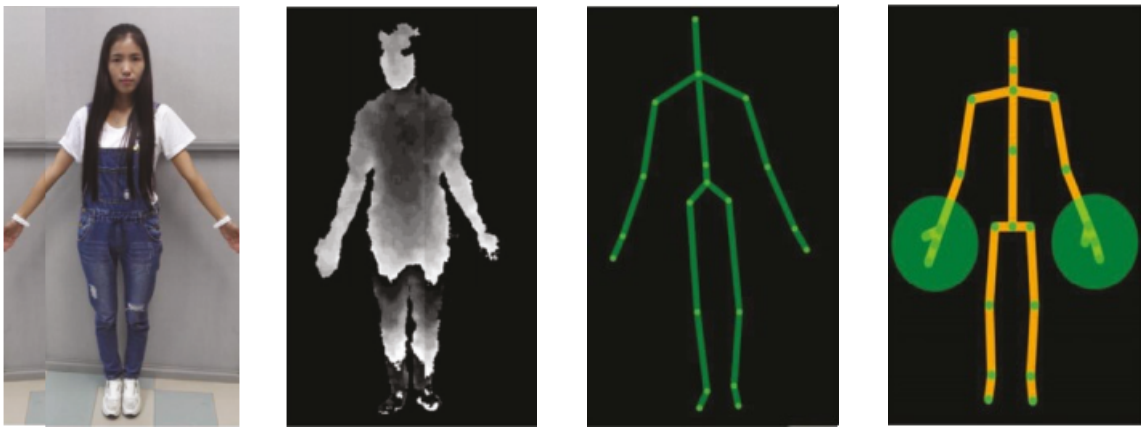
Figure 1: Research trajectory of activity representation approaches

only one subject without any human-object or human-human interactions. A HAR system should recognize as "playing football" rather than as "a group of people running". It is obvious to have occasions of occlusions in crowded places, low quality videos from cameras installed in high places where target person is not clear.

Significance of RGB-D data in HAR: A RGB-D image is simply a combination of a RGB image and its corresponding depth image. A depth image is an image channel in which each pixel relates to a distance between the image plane and the corresponding object in the RGB image. The depth images provide additional information in comparison with the conventional RGB images. Kinect provides a powerful skeleton tracking algorithm, which outputs 20 3D human joint positions for each frame in real-time. Since the movements of the human skeleton can distinguish many actions. The RGB image, depth map, skeleton captured using Kinect v1 and Kinect v2 are presented in Figure 2.

1.2 Human Action Datasets

Publicly available RGBD datasets can at the most basic level, remove the need to repeat the data capture and also they provide transparency in the presentation of results and allow scores to be compared on the same data by different researchers, adopting the standard experimental protocols. Datasets play significant role rapid development, objective evaluation and evidence for efficiency methodologies. Hence, the datasets should



(a) RGB image (b) Depth Map (c) Kinect v1 Skeleton (d) Kinect v2 Skeleton

Figure 2: Various informations captured RGB-D sensors

be carefully created or chosen to ensure effective and reliable evaluation of the validity and efficacy of any algorithm under the study. For this task-specific evaluation of algorithm the factors captured by each dataset plays a key role. The significant datasets are presented in table 1.

Table 1: Some significant action datasets

Name of the Dataset	**Level**	**Subjects**	**Actions**	**Videos**	**Examples**
CAD 60/CAD 120	Activity	4	12/10	60120	brushing
MSR Action 3D	Action	10	20	567	high wave
MSR Daily Activity3D	Activity	10	16	320	drink
MSR Gesture 3D	Action	10	12	336	ASL Signs
RGBD-HuDaAct	Action	30	12	1189	get up
UT Kinect-Action	Action	10	10	200	walk
ChaLearn Gesture Challenge	Action	20	850	50000	Mudras
Sheffield Kinect Gesture	Action	6	10	1080	Gestures
BMM Human Action	Action	12	11	660	jumping
M2I	Interaction	22	22	1760	walk
SBU Kinect Interaction	Interaction	7	8	300	approaching
NTU RGB+D	Activity	40	60	56,000	drinking

1.3 Review of Literature

Due to the advancement in computer vision techniques and availability of low cost sensors and high end computational paradigms and also due to their wide applications, Human Action recognition is one of the highly addressed research domain, especially owing to the exploration of depth information. In the recent five years hundreds of problems have been addressed and equal number of novel methods have been developed. A brief review on major categories of works are presented in the following sections.

Figure 3: Samples of MSR Action 3D dataset

RGB based Video Analytics for HAR: One of the challenge normally faced during recognition is extracting the frames from the video sequence, which belong to only the required action and ensuring the continuity of the action along the non uniform time is another challenge. Hence, these methodologies can be classified into space-time, stochastic, rule-based, and shape-based approaches. A brief review of significant works are presented below.

Wang *et al.* proposed a global spatial-temporal three stream CNNs architecture for action feature extraction, comprising of spatial, local temporal and global temporal streams generated respectively from deep learning single frame, optical flow and global accumulated motion features in the form of a new formulation named Motion Stacked Difference Image (MSDI). Hao *et al.* [10] proposed a method which leverages the popular visual representation method, bag of-visual-words (BoVW)/fisher vector (FV), to represent individual videos in each view. Second, the sparse coding algorithm is utilized to transfer the low-level features of various views into the discriminative and high-level semantics space. Third, they employ the multi-task learning (MTL) approach for joint action modeling and discovery of latent relationship among different action categories. SJ Berlin *et al.* proposed a framework for recognizing human interactions based on deep learning based architecture. The Harris corner points and the histogram form the feature vector of the spatio-temporal volume. The feature vector extraction is restricted to the region of interaction. A stacked auto-encoder configuration is embedded in the deep learning framework used for classification.

Depth based Video Analytics for HAR: A. Jalal *et al.* proposed a HAR framework, where depth maps are analyzed by temporal motion identification method to segment human silhouettes from noisy background and compute depth silhouette area for each activity to track human movements in a scene. C. Chen *et al.* [5] proposed a three stage effective local spatio-temporal descriptor for action recognition from depth video sequences, taking the shape discrimination and action speed variations into account. Depth sequence is divided into temporally overlapping depth segments generating DMMs. A. Jalal *et al.* proposed spatio-temporal features approach to detect, track, and recognize human silhouettes using a sequence of RGB-D images. X. Yang and Y. Tian [32] proposed framework for recognizing human activities from video sequences captured by depth cameras.

1.4 Evaluation Process

It is essential that we adopt a standard metric for presenting our results, in order to facilitate a standard measure of the results achieved with the proposed descriptors and fair comparison of our methods with the existing methods. Classification accuracy is generally stated in terms of overall percentage of correct recognition to the actual samples considered for the experimental trial. It is simply the rate of correct classifications. However, we might consider an independent test or cross-subject test or some variation of cross-validation.

Table 2: Recent/Significant publications on Action Recognition

Author/Year	Key Representation	Classification	Modality/Level	Dataset	Accuracy
Jalal et al.[14] **2017**	Multifeatures extracted from human body silhouettes and joints information	HMM	Depth Modality Interaction Action/Activity	Self Annotated Dataset MSR Daily Activity3D MSR Action3D	71.6% 92.2% 93.1%
Yang et al. [32] **2017**	Low-level polynormal assembled from local neighboring hypersurface normals and are then aggregated by Super Normal Vector	Linear classifier	Depth Modality Interaction Action/Activity Action primitive	MSR Action3D MSR Gesture3D MSR Action Pairs3D MSR Daily Activity3D	93.45% 94.74% 100% 86.25%
Lillo et.al.[18] **2017**	Body poses estimated from RGB-D data	Geometric and motion descriptors - Spatial & Temporal compositions	Depth Modality Action Activities	MSR-Action3D	93.0%
Z.Gao et.al. [8] **2017**	Dense trajectory feature HPM+TM CNN feature	BoW &PCA Collaborative sparse representation leaning model	RGB-Depth Activity	M2I N-UCLA	84.3% 91.5 %
Ibaez et. al[13] **2017**	Body Joints identified by Kinect	Gestures into strings and String Matching using K Means algorithm	RGB Modality Activity	MSRC-12 Kinect	92.00 %
Chen et al [6] **2017**	space-time autocorrelation of gradients (STACOG)	weighted fusion scheme based on ELM	RGB-Depth Activity	MSR Action 3D MSR Gesture 3D	94.87 % 98.50 %
Ijjina et al.[13] **2017**	motion sequence information into CNN	deep learning	RGB-Depth Activity	SBU Kinect	90.98 %
Hu et al., [11] **2017**	joint heterogeneous features learning (JOULE) model	linear SVM classifier	RGB-Depth Action/Activity	MSR Daily Activity CAD-60	95 % 84.1 %

1.5 Motivation and Objectives

Though Human Action Recognition is one of the highly addressed domain by the research community and a significant number of works is being published over the time, it remained to be a interesting and equally challenging problem domain. In spite of the quality research and attaining state of the art results, many problems such as robustness to occlusion and view variance, real time action recognition, recognition of group activities, significant salutation of the complementary features of multi modal approaches are not attained to the fullest of possibility. The availability of multi-model information which are supplement to one another such as RGB information along with depth and skeleton information, has given new dimensions to application based research. More complex and challenging datasets with huge number of samples comprising inter-class and intraclass variations are publicly available for the research community, which facilitates to experiment the models on wide variety of data, leading to development of robust methodologies.

Having carried out detailed review of literature on existing RGBD based HAR works, based on the discerned following actualities of human action recognition. The key objectives planned to achieve in the due course of this research work are identified and enlisted alongside.

- Accumulating the motion information from a video sequence is one of the highly challenging and significant phase in Human Action Recognition. Our goal is to devise a more efficient and simple method to produce a concise representation from depth sequence.

- From the literature it is evident that Wavelet transforms such as dual tree complex wavelets transforms are capable of representing digital signals better than traditional signal transforms like discrete Fourier transform(DFT), discrete cosine transform(DCT) and discrete wavelet transform (DWT). In this context, we intend to explore the possibility using wavelet based features in the frequency domain to build a suitable feature descriptor for extracting the efficient features from the action cues.

- The design of most of the well known local descriptors are based on image gradient, as image gradient is the essential characteristic in the spatial domain. Hence we propose to devise a more accurate feature representation method, exploiting the gradient and orientation information in order to effectively represent the features from the action cues, to develop an effective Human Action Recognition system.

- It is well proved that the partial sum of Tylor Series Expansion (TSE), computed with finite number of terms is capable of representing the local features. Research has established the fact that Riesz filter extracts prominent features due to its ability to register the localized variations in texture and simplicity in modeling. Similarly, 1D intrinsic signals such as lines and edges of an image can be well captured by the first order Riesz filters. In this context, we investigate the possibility of building a multi representative feature descriptor using Raiz and Tylor representations and develop efficient representations for robust Human Action Recognition.

- Deep neural networks such as Convolutional Neural Networks has become a commonly used method for learning image contents. The composition of multiple level of nonlinear operations are capable of determining a complicated decision function from a available data, in the case of such architectures. Hence, we intend to explore the possibility using a deep learning model on RGBD action videos.

2 UDTCWT based Descriptor

In order to shuffle off the burden of accessing huge video files and high computationally intensive 3-layer RGB video frame processing, we need to explore the possibilities of action recognition solely based on depth frames. Further, to extract the textural features over the video frames, we need an efficient and concise representation of the entire depth frame sequence. The proposed work aims to explore depth motion map concept to efficiently represent and classify human actions. The work date backs to the pioneering work carried out by Bobik *et al.* where an accumulated foreground regions of human to track the shape changes using motion history images is addressed. With this backdrop, we propose a compact and computationally efficient DMM representation, namely Stridden Depth Motion Map (S-DMM), which hastily generates an accumulated dense structure striding 4 frames of the video sequence at a time. This completes the DMM generation four times faster than the existing approach presented in [2]. Second, we explore a Undecimated Dual Tree Complex Wavelet Transform (UDTCWT) based feature descriptor for extracting the wavelet features from our S-DMMs. Third, using a Sequential Extreme Learning Machine (S-ELM) for classification, we compare our method with the state-of-the-art methods considering standard datasets.

2.1 Proposed Methodology

The proposed method has three phases namely, S-DMM based depth video representation, UDTCWT based feature extraction followed by classification using S-ELM technique. Different phases of our methodology such as S-DMM generation, UDTCWT feature representation, dimensionality reduction and classification are briefed below. The work flow is pictorially presented in Figure 4.

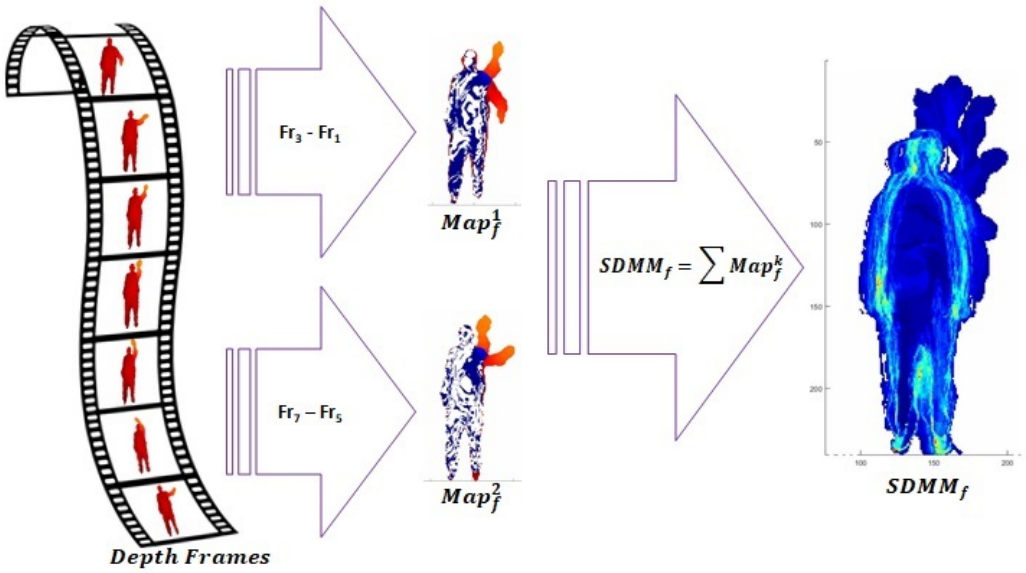


Figure 4: Proposed methodology for generating Stridden Depth Motion Map from depth sequences.

Stridden Depth Motion Maps: Aiming for computation efficiency and more compact representation, we traverse the depth video in strides (steps) of 4 frames per iteration during DMM generation. In addition, we compute the frame variation at an interval of two frames rather than subtracting the consecutive frames to find the energy difference. The summation of absolute differences between two alternative frames across the video sequence yields a single S-DMM. The process is summarized in equation (1).

$$SDMM_{\{f,s,t\}} = \sum_{k=3,k=k+4}^{N} \left| D^{k}_{\{f,s,t\}} - D^{k-2}_{\{f,s,t\}} \right| \tag{1}$$

where f, s, t are *front*, *side* and *top* projections respectively
$D^{k}_{\{f,s,t\}}$ is the k-th frame under the projection view
N is the number of frames in the depth sequence
$SDMM$ is the generated Stridden Depth Motion Map.

In addition to the *front (f)* view calculated as described above, two additional 2D projection views, *side(s)* and *top(t)* are computed from the depth video frames using (2) and (3) respectively.

$$D^{s}_{ik} = \begin{cases} j, & \text{if } D^{f}_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$D^{t}_{kj} = \begin{cases} i, & \text{if } D^{f}_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $i = 1, 2...M, j = 1, 2...N$ and $k = 1, 2...L$
L being the maximum depth value across *front* frames
D^{f} is the current *front* frame of size *MxN*
D^{s} and D^{t} are the generated *side* and *top* projections.

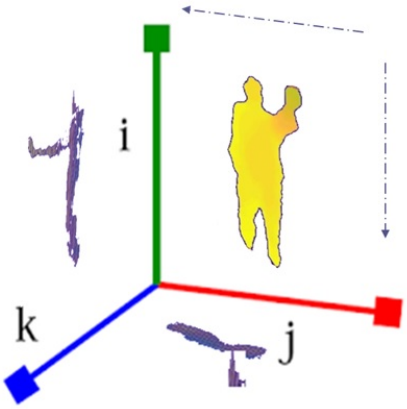


Figure 5: Producing side and top views from the frontal view

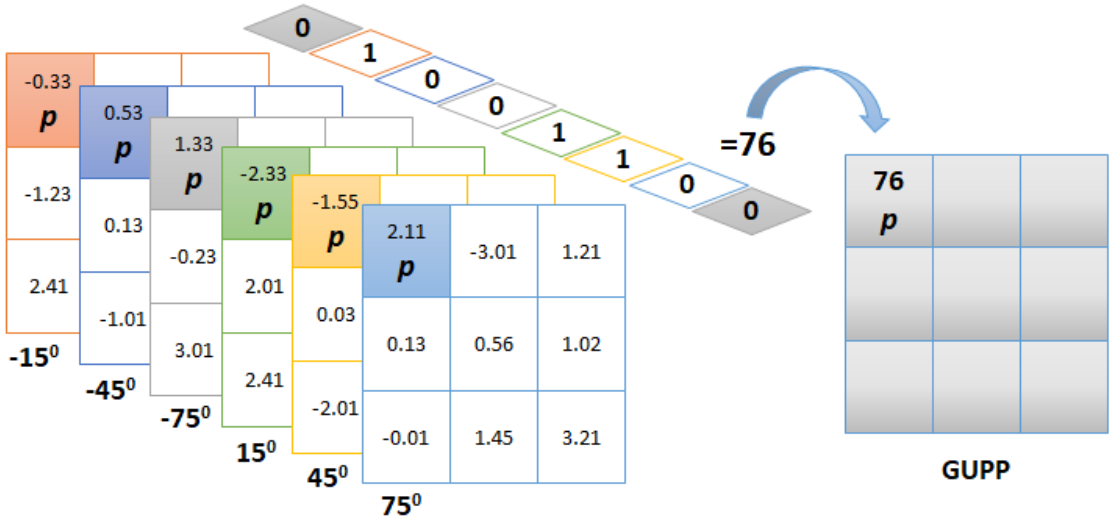


Figure 6: Computing GUPP value at location p from complex coefficient images at six different orientations.

Undecimated Dual Tree Complex Wavelet Transform: During the second phase, each of the S-DMMs generated as described above are scanned with an Undecimated Dual Tree Complex Wavelet Transform (UDTCWT) function to produce an efficient feature descriptor. The traditional wavelet transforms such as DWT, DTCWT are having certain limitations when applied in digital image processing[24]. For example, the wavelet function produced from 2D DWT is not suitable for extracting diagonal edge details of an image due to its checker board pattern spectrum, whereas Dual Tree Complex Wavelet Transform (DTCWT) is free from such limitations. DTCWT wavelet coefficients are invariant to signal shift, free from oscillations around signal discontinuities such as edges and moreover do not suffer from aliasing.

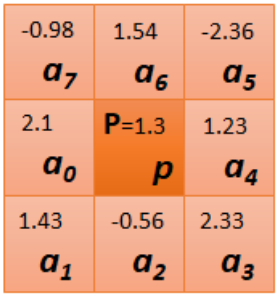


Figure 7: Eight neighbor locations for computation of LUPP value at p.

An undecimated version of the DTCWT is more robust than DWT and DTCWT as downsampling is not employed, which helps in incorporating perfect one-to-one relationship between the co-located coefficients extracted at different scales. This helps to attain a flawless shift invariance in UDTCWT.

In our work, we form an higher dimensional global descriptor using UDTCWT similar to the method presented in [24]. This is achieved in four stages. At first, we compute UDTCWT at 6 different orientations (-15^0, -45^0, -75^0, 15^0, 45^0, 75^0) and 4 scales

leading to 24 complex coefficient images. We separate the real and imaginary part of each coefficient image producing 48 complex coefficient images in total. During the second stage, we generate 8 Global UDTCWT Phase Pattern (GUPP) images from the above said 48 complex coefficient images. In this process, the values of 6 orientation coefficient images at each location (x,y) are concatenated forming a 6 element vector. The elements in vector are binarised (0 if value is greater than or equal to 0, 1 otherwise), to get a 6 bit vector which is in turn 0 padded at the two ends to form a vector byte. The value of this byte is the GUPP image value at (x,y). The process is shown in Figure 6. GUPP image for a given scale is computed by computing the values in the similar manner for all (x,y) locations.

In the third stage, for each of the 48 complex coefficient images Local UDTCWT Phase Pattern (LUPP) images are generated. The LUPP image value at location (x,y) is computed from its eight neighbor locations as shown in Figure 7, forming a 8 bit vector. LUPP value at location p is computed using equation 4 as shown below,

$$LUPP(p) = \sum_{j=0}^{7} sgval(a_j * P) * 2^j \tag{4}$$

where sgval(product)=0 if product is +ve, =1 if product is -ve.
P is the value of UDTCWT at the center pixel p.
a_js are values of UDTCWT at the eight neighbors.

Finally, we divide GUPPs and LUPPs into blocks, compute spatial histogram in each block and concatenate these histograms to form a global descriptor for the given S-DMM. We as well extract similar features from *side* and *top* projections of S-DMMs.

Dimensionality Reduction: Due to consecutive concatenation of spatial histogram of each blocks and concatenation of similar vectors from three projected DMMs, the dimension of feature vector turns to be very high. We employ Principal Component Analysis (PCA) to reduce the dimension of feature vector, retaining the most discriminating features.

Classification using S-ELM: We use Sequential Extreme Learning Machine for classification. S-ELM is capable of randomly choosing the input weights and analytically determining the output weights of SLFNs.

2.2 Experimental Analysis

Experiments were conducted on two standard datasets namely, MSR Action 3D dataset and DHA dataset.For comparison with the standard published results, we have examined our method under two different experimental settings available in the literature[2]. To evaluate our method on DHA dataset, we employ Leave One Subject Out (LOSO) testing protocol. Under setting 1 of MSR Action 3D dataset, our method achieves 95.75% of average accuracy. The results are tabulated in Table 3. Under setting 2 with cross subject test, experimental results are obtained and the results are tabulated in Table 4 along with other existing methods. The comparative results on DHA dataset are presented in Table 5. The confusion matrix is presented in Figure 10, that demonstrates the class-wise recognition accuracies of MSR Action 3D dataset under setting 2 and here 7 out of 13 *Draw x* actions are wrongly classified as *Draw tick*.

Table 3: Average recognition accuracies (%) under Cross Subject tests on fixed subsets on MSR Action 3D dataset

Method		Action Subsets			Average (%)
		AS1	AS2	AS3	
Li et al.[17]		72.9	71.9	79.2	74.7
DMM-HOG [34]		96.2	84.1	94.6	91.6
Chen et al. [3]		96.2	83.2	92.0	90.5
HOJ3D [31]		88.0	85.5	63.6	79.0
STOP [28]		91.7	72.2	**98.6**	87.5
DMM-LBP [2]		**98.1**	92.0	94.6	94.9
Proposed	DMM-UDTCWT	95.6	**93.82**	96.6	95.34
	Stridden DMM-UDTCWT	96.52	**93.82**	96.92	**95.75**

Table 4: Recognition accuracies under setting 2 on MSR Action 3D dataset .

Method	Accuracy (%)
DMM-HOG[34]	85.5
ROP [29]	86.5
HON4D [23]	88.9
DMM-LBP [2]	91.9
DMM-UDTCWT	92.67
Stridden DMM-UDTCWT	**93.41**

3 Complex Gradient based Descriptor

The basic idea of design of these descriptors is to transform the image regions using a class of filters and then computing a invariant distinctive features, emphasizing the image properties like pixel intensities, color, texture, edges, etc. Since, local descriptors are at the vital course of many computer vision tasks, ample of descriptor tools have been proposed in the literature to improvise the efficiency and accuracy of the image processing algorithms. Ojala *et al.* have introduced LBP, as a gray scale invariant local binary descriptor, defined on a 3×3 neighbourhood. The central pixel of the neighbourhood is obtained by decimalizing the local binary pattern obtained by the differences of the sample pixel values from the central value. SIFT descriptor, as introduced by Lowe is created by first computing the gradient magnitude and orientation at each image sample point in a region around a keypoint location. These keypoints are weighted by a Gaussian window and then accumulated into orientation histograms summarizing the contents over 4×4 subregions. Soon after this, Dalal and Triggs have introduced histogram oriented gradient descriptor HOG.

3.1 Proposed Methodology

The architecture and design of the proposed descriptor is simple to implement. The image gradient is a robust property of gray scale images which exhibits important features of the image such as, illumination changes, the amount of change and the direction through which pixel value varies. We have mapped each pixel into the complex domain of gradient magnitude and direction.

Table 5: Average recognition accuracies under Leave One Subject Out test on DHA dataset

Method	Accuracy (%)
D-STV/ASM [19]	86.8
SDM-BSM [21]	89.5
D-DMHI-PHOG [23]	92.4
D-STV [9]	86.8
DMM-UDTCWT	94.2
Stridden DMM-UDTCWT	**94.6**

Complex gradient function: In this section, we present the theoretical explanation of the descriptor. Considering the gray scale image as a real valued 2D function $f(x, y)$, image gradient at (x, y), $\nabla f(x, y)$, is represented as vector given by,

$$\nabla(f(x,y)) = [f_x, f_y]^T(x,y) = \Big[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial x}\Big]^T(x,y), \tag{5}$$

whose magnitude and direction at (x, y) are,

$$mag(\nabla f)(x,y) = \sqrt{f_x^2(x,y) + f_y^2(x,y)}, \tag{6}$$

$$dir(\nabla f)(x,y) = tan^{-1}\Big(\frac{f_y}{f_x}\Big). \tag{7}$$

Thus, each pixel is mapped on to the gradient field $\mathcal{C}(\omega, \alpha)$ such that,

$$f(x,y) \rightarrow (mag(\nabla f)(x,y), dir(\nabla f)(x,y)) \rightarrow (\omega, \alpha) \in \mathcal{C} \tag{8}$$

Existing gradient based descriptors have well utilised these information to design the feature extraction tools. However we capture each and every localalised information at a pixel (x, y) by considering each (ω, α) as a complex representation of $f(x, y)$ and formulate our descriptor using the magnitude and amplitude of (ω, α). We map the plane $\mathcal{C}$ to a complex gradient domain $\mathcal{CGF}(\xi(x,y), \vartheta(x,y))$, which is obtained by the complex gradient function defined below.

$$\xi(x,y) = \sqrt{\omega^2(x,y) + \alpha^2(x,y)}, \tag{9}$$

$$\vartheta(x,y) = tan^{-1}\Big(\frac{\alpha(x,y)}{\omega(x,y)}\Big), \tag{10}$$

Thus, input image is described with $\mathcal{CGF}_{Mag}(\xi, \vartheta)$, the field of magnitudes and $\mathcal{CGF}_{Amp}(\xi, \vartheta)$, the field of directions.

Mapping of $\mathcal{C}$ onto $\mathcal{CGF}$ has mainly two advantages. Firstly, the combined effect of magnitude and orientation can be well described in $\mathcal{CGF}$ domain. Secondly, the amalgamated structural information of both gradient magnitude and directions is easily extracted from the two fields $\mathcal{CGF}_{Mag}$ and $\mathcal{CGF}_{Amp}$.

$\xi(x, y)$ and $\vartheta(x, y)$ give normalized impact of orientation and magnitude of image gradient at the pixel (x, y) and will describe the input pixel in terms of its complex gradient function ($\mathcal{CGF}$). Thus, $\mathcal{CGF}(x, y)$ describes each input pixel as a two dimensional vector whose components comprise of both the information of ω and α. Hence, compared to existing LFD's, CGF is more robust in identifying the predominant local features of image texture. Robustness of CGF has been justified in the subsequent Sections wherein, we present the empirical evidences by deploying CGF for different applications.

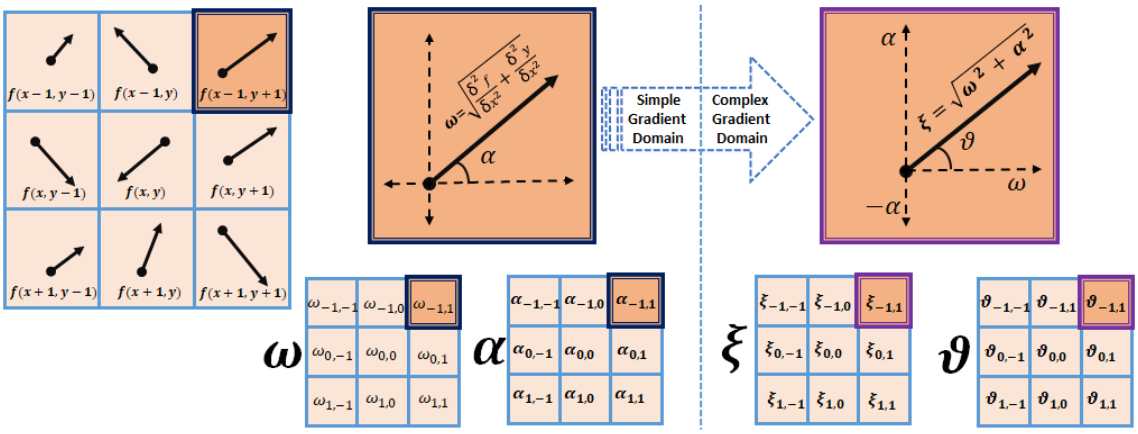


Figure 8: Pictorial representation of proposed approach for computing CGF

Compact Representation using SDMM: In order to produce a computationally efficient SDMM, the depth video is traversed with a striding (in steps) of four frames per iteration and the variations among the frames are calculated for alternate frames. Thus, here the frame variation is the absolute difference between two alternate frames with a stride of 4 frames. These absolute differences are summed up along the video frames, producing a SDMM.

ELM based Classification: We use the CGF filter to obtain the CGF representation as formulated in Equation 9. Blockwise histogram over the magnitude factor of the CGF is computed and concatenated to produce the final feature vector. A kernal based extreme learning machine is used for classification.

3.2 Experimental Results

The CGF descriptor based methodology achieved significantly better results than many of the existing methods available in the literature. The experiments using the CGF framework on subsets AS1, AS2 and AS3 of MSR Action3D dataset under TestOne and TestTwo testing strategies demonstrate competitive results on par with the results presented in existing literature, as presented on Table 6 and Table 7 respectively.

Our experiments on MSR Action3D shows an average accuracy of 96.39% under cross subject testing strategy of setting1, as presented in Table 8 whereas, Table 9 presents the cross subject test results on the same dataset under setting2 demonstrating significant improvements compared to the available results. Referring back to the works such as [27, 34], it is observant that inter-class similarity among the samples belonging to *Draw tick*, *Draw x* and *Draw circle* is very high, also the samples belonging to these classes are mostly misclassified in the above cited existing works. In our experimentation, it is evident that six samples belonging to *Draw x* are misclassified as *Draw tick* class.

However, owing to the robustness of proposed method, only one sample belonging to *Draw x* is wrongly classified into *Draw circle* class. The proposed methodology investigates the applicability of newly devised hand crafted CGF descriptor for human action recognition from depth videos. Considering the wide exploration of neural network based models on HAR domain, we present a comparison of our results with some existing NN model based results on MSR Action3D dataset in Table 10.

Table 6: Comparison of recognition rates(%) attained under Fixed Test-Test One for MSR Action3D Dataset

Method	Subsets of Classes			Avg.(%)
	AS1	AS2	AS3	
Li et al.[17]	89.5	89	96.3	91.6
DMM-HOG [34]	97.3	92.2	98	95.8
Chen et al.[2, 3]	97.3	96.1	98.7	97.4
HOJ3D [31]	98.5	96.7	93.5	96.2
Yang et al.[33]	94.7	95.4	97.3	95.8
STOP [28]	98.2	94.8	97.4	96.8
Proposed	**97.3**	**97.3**	**100**	**98.2**

Table 7: Comparison of recognition rates(%) attained under Fixed Test-Test Two for MSR Action3D Dataset

Method	Subsets of Classes			Avg.(%)
	AS1	AS2	AS3	
Li et al.[17]	93.4	92.9	96.3	94.2
DMM-HOG [34]	98.7	94.7	98.7	97.4
Chen et al.[2, 3]	98.6	98.7	100	99.1
HOJ3D [31]	98.6	97.2	94.9	97.2
Yang et al.[33]	97.3	98.7	97.3	97.8
STOP [28]	99.1	97	98.7	98.3
Proposed	**100**	**98.7**	**100**	**99.5**

Table 8: Comparison of recognition rates(%) attained under Fixed Test-Cross Subject Test for MSR Action3D Dataset

Method	Subsets of Classes			Avg.(%)
	AS1	AS2	AS3	
Li et al.[17]	72.9	71.9	79.2	74.7
DMM-HOG [34]	96.2	84.1	94.6	91.6
Chen et al.[2]	96.2	83.2	92.0	90.5
HOJ3D [31]	88.0	85.5	63.6	79.0
STOP [28]	91.7	72.2	98.6	87.5
DMM-LBP [2]	98.1	92.0	94.6	94.9
SDMM-UDTCWT [27]	96.52	93.82	96.92	95.75
Proposed	**97.22**	93.75	**98.20**	**96.39**

Table 9: Recognition accuracies achieved with setting 2

Method	Accuracy(%)
DMMHOG[34]	85.5
ROP [29]	86.5
HON4D [23]	88.9
DMM-LBP [2]	91.9
SDMM-UDTCWT [27]	93.41
Proposed	**95.24**

Table 10: Comparative result analysis achieved with some existing NN models, based on the performance on MSRAction3D dataset

Method	Subsets of classes			Avg.(%)
	AS1	AS2	AS3	
3D CNN[22]	86.79	76.11	89.29	84.07
Zho et al. [35]	95.89	87.68	97.32	94.15
H-RNN [7]	93.33	94.64	95.50	94.49
Proposed	**97.22**	93.75	**98.20**	**96.39**

Proposed CGF descriptor enhances the utilization of image gradient with the combined impact of magnitude and direction. It can be further explored to other applications of image processing such as, face recognition, multi-modal bio-metrics, object recognition and human action analysis.

4 Stacked Filter Bank based Descriptor

The proposed work explores Stridden Depth Motion Map (SDMM) concept for effectively and concisely representing the action cues from depth sequences generating an accumulated dense structure. Motivated by the success of UDTCWT based feature descriptor presented in Shekar et.al. [27], we propose a new Stacked Filter Bank(SFB) based descriptor which results in a significantly (roughly $\frac{1}{4}^{th}$ times) smaller and more discriminative feature vector. Motivated by the promising results in Shekar et.al. [25, 26] which exploit TSE and Riesz filters for iris recognition, we propose a pool of filters based partial differentials as the first and higher derivatives which are good example to intricate local features.

4.1 Proposed Methodology

The methodology is realized in three phases. Initially, a compact SDMM representation with hasty striding is computed and during the second phase, the SDMM is processed with the SFB, which stacks up the responses from TSE filter, Riesz Filter and gradient filter. The following subsections present these phases in detail.

Compact Representation using SDMM: In order to produce a computationally efficient SDMM, the depth video is traversed with a striding (in steps) of four frames per iteration and the variations among the frames are calculated for alternate frames. Thus, here the frame variation is the absolute difference between two alternate frames with a stride of 4 frames. These absolute differences are summed up along the video frames, producing a SDMM.

Stacked Filter Bank based descriptor: In order to extract the features effectively, we propose a stack of three filters namely TSE, Riesz and gradient filters. Each of these filters are capable of yielding two feature-rich responses of its own kind. These feature-rich representations are stacked to give a common response which is intern used to form the feature descriptor.

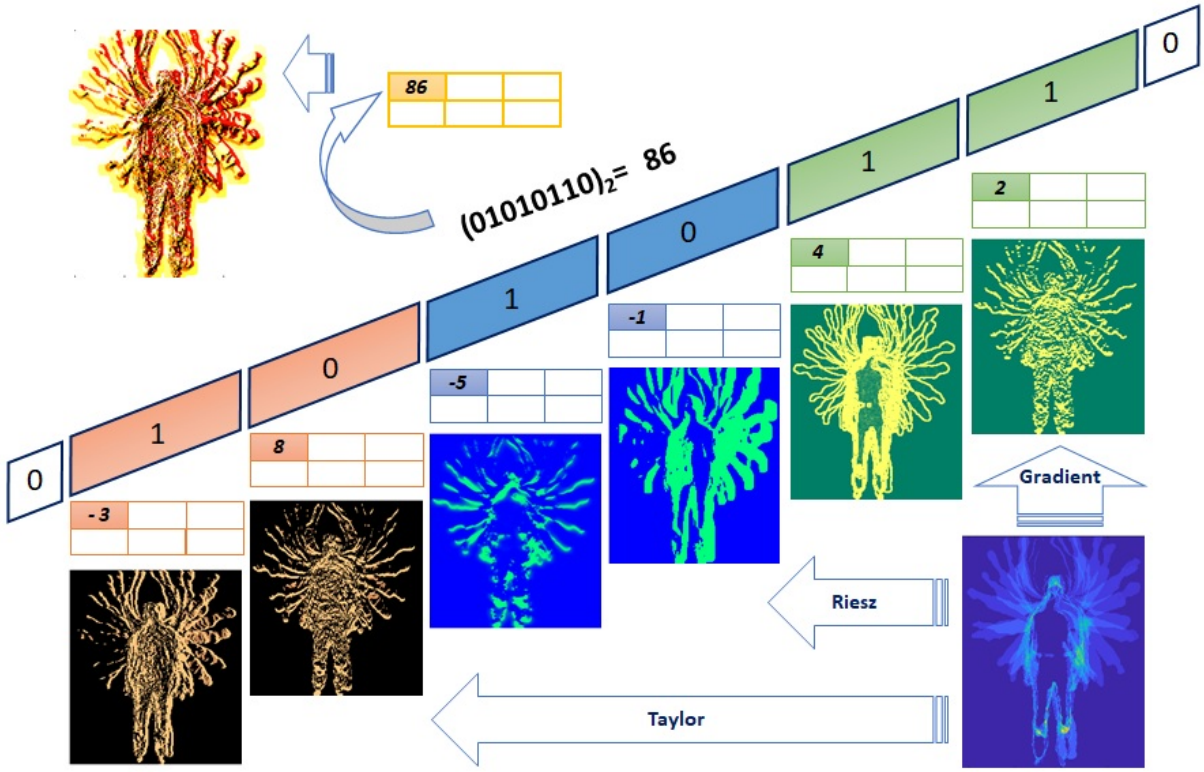


Figure 9: Methodology for obtaining Stacked Filter Bank based Response from SDMM

TSE Filter: Basically, Taylor series expansion of a function $f(x)$ at a point $x = a$ estimates the function in a small neighborhood of the point a. The expression for infinite Taylor series expansion of a regular function $f(x)$ at $x = a$ is,

$$f(x) \equiv \sum_{n=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n \tag{11}$$

If one is able to successfully find the functional value and the values of its derivatives at $x = a$, then one can also estimate the functional values at all the points in the neighborhood of a. It can also be written as,

$$f(x) \equiv \sum_{n=0}^{N} \frac{f^n(a)}{n!}(x-a)^n + R_n \tag{12}$$

where R_n is the Taylor's remainder.

As R_n tends to zero as $n \to \infty$, the partial sum of TSE computed with N finite number of terms in a small neighborhood gives approximation for the function $f(x)$. This property of TSE helps in effectively capturing the local region features from SDMM.

Riesz Filter: The extensions of Hermite functions into d dimensional Euclidean space, are considered as Riesz transformations. Riesz transformation R of a d dimensional function $f(x) \in L_2(R^d)$ is a vector signal which transforms a signal from $L_2(R^d)$ to $L_2^d(R^d)$, where $x = (x_1, x_2, ..., x_d)$, and is given by

$$Rf(x) = ((h_1 * f)(x), (h_2 * f)(x), ..., (h_d * f)(x)) \tag{13}$$

where each Riesz kernel h_i is given by

$$h_i = \frac{\Gamma\frac{(d+1)}{2}}{\pi^{(d+1)/2}} \frac{x_i}{\|x\|^{(d+1)}} \tag{14}$$

where Γz represents gamma integral value of any real number z and $\|.\|$ is the Euclidean norm. 2D Riesz kernels are achieved by substituting $d = 2$ in the above equation.

$$h_x = \frac{1}{2\pi} \frac{x}{\|x\|^3} \ , \quad h_y = \frac{1}{2\pi} \frac{y}{\|x\|^3} \tag{15}$$

Gradient based Filter: Considering the gray scale image as a real valued 2D function $f(x, y)$, image gradient at (x, y), $\nabla f(x, y)$, is represented as vector given by,

$$\nabla(f(x, y)) = [f_x, f_y]^T(x, y) = \Big[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial x}\Big]^T(x, y), \tag{16}$$

whose magnitude and direction at (x, y) are,

$$mag(\nabla f)(x, y) = \sqrt{f_x^2(x, y) + f_y^2(x, y)}, \tag{17}$$

$$dir(\nabla f)(x, y) = tan^{-1}\Big(\frac{f_y}{f_x}\Big). \tag{18}$$

$$f(x, y) \rightarrow (mag(\nabla f)(x, y), dir(\nabla f)(x, y)) \rightarrow (\omega, \alpha) \in \mathcal{C}. \tag{19}$$

Thus, each pixel is mapped on to the gradient field $\mathcal{C}(\omega, \alpha)$, and we utilize these information into our SFB to design a feature representation.

Filter Bank based descriptor: Each of the above filters map a single pixel into two real responses, resulting six responses in total, which are stacked together and a common response image(CRI) is computed. In the due course, the pixel values of six responses at each (x,y) locations are treated as common 6 element vector. We intend to convert this to a representative eight bit binary vector. In order to achieve this, the elements in the vector are first binarized on zero crossings (treated as 0 if the value is greater than or equal to 0, treated as 1 otherwise), generating a six bit vector. Then, this vector is padded with a 0 at both the ends to get a byte vector. The decimal equivalent value of this byte is computed and this decimal value is the CRI point value at (x,y). The process is demonstrated in Fig. 9. To best utilize the individual response image (IRI) based features, we also compute a IRI for each of the response image. The IRI value at a point (x,y) is calculated based on the values of its eight neighboring pixel locations forming a 8 bit vector. The computation of IRI value at location p is based on the (20) as shown below,

$$IRI\big(p\big) = \sum_{j=0}^{7} sgval(a_j * P) * 2^j \tag{20}$$

where sgval (product) is equal to 1 if the product is -ve value, and equal to 0 if the product is +ve value. P is the Filter Response value received at the center pixel p. a_js are values of Filter Response at the eight neighbors.

Table 11: Cross Subject test based average recognition rates(%) on fixed subsets of MSR Action 3D dataset

Method	Action Subsets			Avg.(%)
	AS1	AS2	AS3	
Li et al.[17]	72.9	71.9	79.2	74.7
DMM-HOG [34]	96.2	84.1	94.6	91.6
Chen et al. [3]	96.2	83.2	92.0	90.5
HOJ3D [31]	88.0	85.5	63.6	79.0
STOP [28]	91.7	72.2	98.6	87.5
DMM-LBP [2]	**98.1**	92.0	94.6	94.9
SDMM-UDTCWT [27]	96.52	**93.82**	96.92	95.75
Proposed	97.22	92.85	**99.09**	**96.38**

Table 12: Recognition accuracies under setting 2 on MSR Action 3D dataset

Method	Accuracy (%)
DMM-HOG[34]	85.5
ROP [29]	86.5
HON4D [23]	88.9
DMM-LBP [2]	91.9
SDMM-UDTCWT [27]	93.41
Proposed	**96.70**

Finally, the CRIs and IRIs are divided into blocks of size 32x32, the spatial histogram in each block is computed and these histograms are concatenated together. These concatenated spatial histogram forms the global descriptor for the given SDMM. However, our descriptor is significantly smaller in size compared to the methods proposed therein and hence computationally less expensive.

Classification using CRC: The proposed method demonstrates significant improvements over many of the existing published results. Our method achieves 96.38% of average accuracy under setting 1 of MSR Action 3D dataset, which is comparatively higher as presented in Table 11. The cross subject test under setting 2 of the dataset poses significant improvements over existing results and is presented in Table 9 along with a comparison of results obtained by existing methods. We observe that six of the *Draw x* action samples are wrongly classified into *Draw tick* class. However, only one *Draw x* sample is misclassified as *Draw circle*. This is due to the strong interclass similarity among these three action sequences.

Comparison with Neural network based models: Though this work proposes a hand crafted Filter Bank based descriptor which extracts the local features for HAR on depth sequences, motivated with the emerging trend of neural network and deep learning models we compare our results with the published results of some existing NN models as presented in Table 13. It is observed that our methodology achieves state-of-the-art results on par with the published results. It is also observed that under setting 2, our

Table 13: Comparison of the results on MSR Action 3D dataset using our model with some existing neural network models

Method	Action Subsets			Avg.(%)
	AS1	AS2	AS3	
3D CNN[22]	86.79	76.11	89.29	84.07
Zho et al. [35]	95.89	87.68	97.32	94.15
H-RNN [7]	93.33	94.64	95.50	94.49
Proposed	**97.22**	92.85	**99.09**	**96.38**

methodology achieves 96.70% of average accuracy whereas the work by Zhao et al.[35] presents an average accuracy of 94.15%.

This work demonstrates a more efficient human action recognition method using SFB based Descriptor. Our experiments exhibit better results, with promising class-wise recognition improvement compared to state-of-the-art methods. Moreover, the proposed method generates significantly compact feature vector, with less computation compared to the UDTCWT feature descriptor. This work demonstrates the applicability of proposed feature descriptor for action recognition in a single subject experimental scenario, which can be extended to multi-subject scenarios as a future work.

5 Neural Network based Model

Recently, researchers have directed remarkable progression to the application of deep learning in human action recognition. The classical image processing and machine learning techniques which use the handcrafted features are not robust and efficient, given any input data. Instead, those methods are well suited only on the data acquired in controlled environment. The significant feature of deep learning, their capability of automated feature learning from images makes gives them a upper-hand. Various nueral network based model and deep learning models have been proven efficient to extract action features from RGB data, depth data, and skeleton sequences from RGB-Depth video sequences.

5.1 Proposed Methodology

A simple way of achieving better performance in a neural networks is by making it size. We tend to increase both depth, the number of levels and width, the number of units of each layer of neural network. However, this leads to larger number of parameters, which in turn makes the network to over-fit in the cases of smaller training sets. Hence, an optimally designed high quality data training set becomes essential to train such networks. At the same time, these approaches become expensive to implement as it demands increased use of computational resources.

Inception v1 which is known as GoogLeNet has 9 inception modules stacked linearly, with 22 layer deep and 5 pooling layers. The last inception module uses a global average pooling.However, this model suffers from vanishing gradient problem.In order to avoid the loss in the middle part, later two auxiliary classifiers were introduced, in the form of softmax to the outputs of two of the inception modules, and computed an auxiliary loss over the same labels. Later, factorization techniques were used to improvise the efficiency of the

model. A 5x5 filters were factorized into two 3x3 filters, which boosted the performance.A convolution of filter size nxn is factorized into a combination of 1xn and nx1 convolutions. Note that, 3x3 convolution is yielding combined result of a 1x3 convolution followed by a 3x1 convolution. This 33% more cheaper than original 3x3 convolution. Deepening the module would lead to excessive reduction in dimensions, and subsequently to information loss, hence filter banks were widened.

It was found that auxiliary classifiers are not contributing for convergence at the beginning of the training process, when training accuracies are near its saturation the network with the auxiliary branches perform remarkably better. Hence, RMSProp Optimizer was incorporated into the model and 7x7 factorized convolutions are used. Apart from the auxiliary classifiers, label smoothing, which prevents over-fitting, is also incorporated. In brief, inception v2 model adopts a factorized version of traditional 7x7 convolution into three 3x3 convolutions and inception v3 adds Batch Norm-auxiliary.

Transfer Learning: Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems. In our experimentation we use inception v3 with pre-trained on ImageNet and transfer learning to CAD-60.

5.2 Experimental Analysis

The basic idea is to convert the video and classification labels in the data set into images (video frames) and their corresponding classification labels, and then use the CNN network to train and test the images to convert the video classification problem into a graphics classification problem. We adopt leave one subject out test strategy for our experimentation on CAD-60 dataset and achieve results on par with existing results as tabulated in Table 14. The specific steps adopted during the experimentation is as follows,

(1) Cut out video frames (jpegs) for each video (training and test videos) at a certain FPS and save them as training set and test set, and use the classification performance of the image as the classification performance of the corresponding video: train set has 1 , 788, 425 frames of images, the test set has 697, 865 frames of images

(2) Select a pre-trained CNN network architecture and initial weights, and transfer learning to CAD-60 dataset, we use inception v3 with pre-trained on ImageNet.

(3) Retrain some layers of the CNN network with train set to obtain the model

(4) After the training is completed, load the model to check and verify all the video frames in the test set, and obtain the top1 accuracy rate and top5 accuracy rate output on the full test set.

In our methodology, raw video frames are extracted and fed to the model. In this work we use a CNN i.e Inception V3, pre-trained on ImageNet. We use transfer learning to retrain Inception v3 on our RGBD data. More specifically, we conducted experimentation on CAD-60 dataset and achieved results on par with existing models under the standard testing protocols.

Table 14: Average recognition accuracies under Leave One Subject Out test on CAD 60 Dataset

Method	Accuracy (%)
Object Affordance [16]	71.40
HON4D [23]	72.70
Actionlet Ensemble [30]	74.70
MSLF[15]	80.36
JOULE-SVM[12]	84.10
Inception V3 +TL	**94.03**

6 Comparative Analysis

To prove the applicability of the proposed methodology on action recognition, we conducted experimentation on standard datasets and present evaluations on a the publicly available depth based action sequence action databases namely MSR Action 3D dataset, DHA Dataset, MSR Gesture 3D dataset and CAD60 datasets. In order to extract significant features from the motion maps we devise three different kinds of feature descriptors and demonstrate their efficiency in terms of classification accuracy. Further, we compare these descriptor based action recognition frameworks on standard datasets, under standard protocols and settings. We adopt the significant datasets and the same protocols as demonstrated in various literature in order to give a fair comparison with evidence. The comparison and analysis of our methods with the existing methods and evaluation of our proposed methodologies in terms of classification accuracy and detailed confusion matrices are presented in Table 15.

6.1 Performance Analysis

Improved Concise Representation : During video based action recognition, representing the action cues in a video sequence in a concise form is a challenging task, as the quality of such representation plays significant variations of the entire video sequence precisely to a feature extractor. In this regard, in spite of several methods available in literature, we establish better and effective motion representation namely Stridden Depth Motion Map. The proposed SDMM is faster and more prominent. Compared to the existing methods of computation proposed by Chen *et al.* [2, 3] our method is 4 times faster as we stride by 4 frames at a time during our method of computation.

Improvement in Recognition Rate : We conducted extensive experimental analysis of our proposed methodologies on standard datasets such as MSR Action3D, MSR Gesture 3D, DHA dataset and CAD 60 dataset. The experimental set up are detailed and results are tabulated in the respective Sections. A consolidated summery of class-wise recognition rates for top 13 action classes of MSR Action3D dataset is tabulated in Table 15 for better understanding. Class-wise recognition rates achieved with our methodologies are presented in the form of confusion matrices in Figures 10. We also present a color coded Class-wise Recognition Rate Comparison in Figure 13.

Table 15: Comparison of recognition accuracies on Action3d dataset

Action Classes	Methodology								
	GLAC[1]	DMM-LBP[2]	MT-DMM[5]	AH-DMM[20]	MT-DMM[4]	WF-2D[6]	**UDTCWT**	**CGF**	**SFB**
Hammer(12)	100%	100%	92%	92%	100%	91.7%	**100%**	**100%**	**100%**
Hand Catch(12)	58.3%	58.4%	75%	83%	75%	66.7%	**100%**	**83.3%**	**100%**
Forward Punch(11)	100%	100%	100%	100%	100%	90.9%	**100%**	**100%**	**100%**
Draw x(13)	53.8%	76.9%	54%	79%	76.92%	46.1%	**46.2%**	**46.2%**	**46.2%**
Draw tick(15)	86.7%	100%	100%	100%	100%	100%	**100%**	**100%**	**100%**
Draw circle(15)	66.7%	80%	93%	73%	86.97%	93.3%	**66.7%**	**93.3%**	**100%**
Hand Clap(15)	100%	100%	100%	93%	100%	100%	**100%**	**100%**	**100%**
Bend(13)	100%	100%	100%	100%	100%	100%	**100%**	**100%**	**86.7%**
Forward Kick(15)	100%	100%	100%	93%	100%	100%	**100%**	**100%**	**100%**
Side Kick(11)	90.9%	72.7%	100%	82%	100%	100%	**100%**	**100%**	**100%**
Tennis Swing(15)	100%	86.6%	100%	100%	100%	100%	**93.3%**	**86.7%**	**100%**
Tennis Serve(15)	86.6%	86.6%	100%	87%	93.33%	100%	**93.3%**	**100%**	**100%**
Pickup throw(14)	100%	92.9%	100%	100%	100%	100%	**92.9%**	**100%**	**100%**

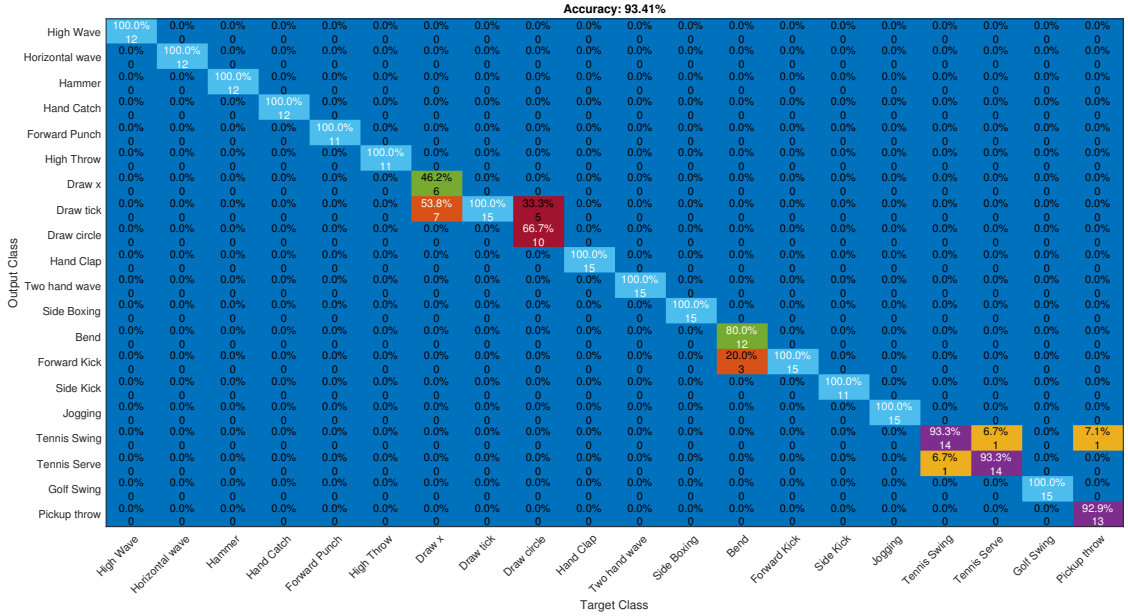


Figure 10: Confusion matrix under setting 2 on MSR Action 3D dataset demonstrating the class-wise recognition accuracy

Inter-class similarity and Self-Occlusion: One of the common challenges in HAR is inter-class similarity. The datasets used in our experimentation pose strong inter-class similarity and self occlusion among multiple action classes in each of the datasets. In spite of these challenges of inter-class similarity, intraclass variation and self occlusions, the proposed descriptors and methodologies perform well and yield comparatively better results than the existing methods.

Self Addressing the Variable Length: Variable length of action video samples is a common problem faced during video based action recognition. This challenge is self addressed by our model. Our method generates a consise Motion Map representation which automatically resolves the issue of varying length of video samples. Irrespective of the video length, our method yields a single image representation namely motion map which brings the entire action cue information of the video. This self solving nature of our method is an added advantage. Also, the size of the descriptor is constant, irrespective of the input length.

7 Conclusion

An accurate and efficient method for human action recognition using UDTCWT based feature descriptor considering newly proposed Depth Motion Map from depth sequences is presented. The proposed Stridden DMMs efficiently and quickly register the action cues and UDTCWT extracts the wavelet features from this compact DMM representation. A Sequential ELM is employed to classify the action sequences. The proposed method is extensively evaluated on two benchmark datasets, under standard protocols presented in the literature. Our experiments exhibited better results when compared to state-of-the-art methods.

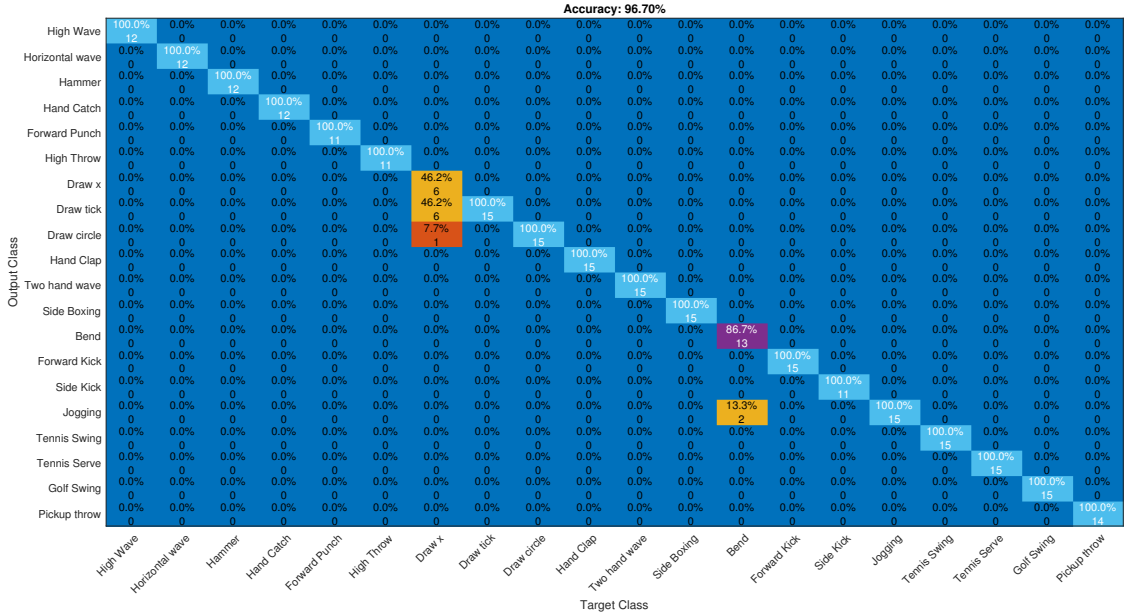


Figure 11: Confusion matrix demonstrating the recognition accuracies for each class under setting 2 on MSR Action 3D dataset

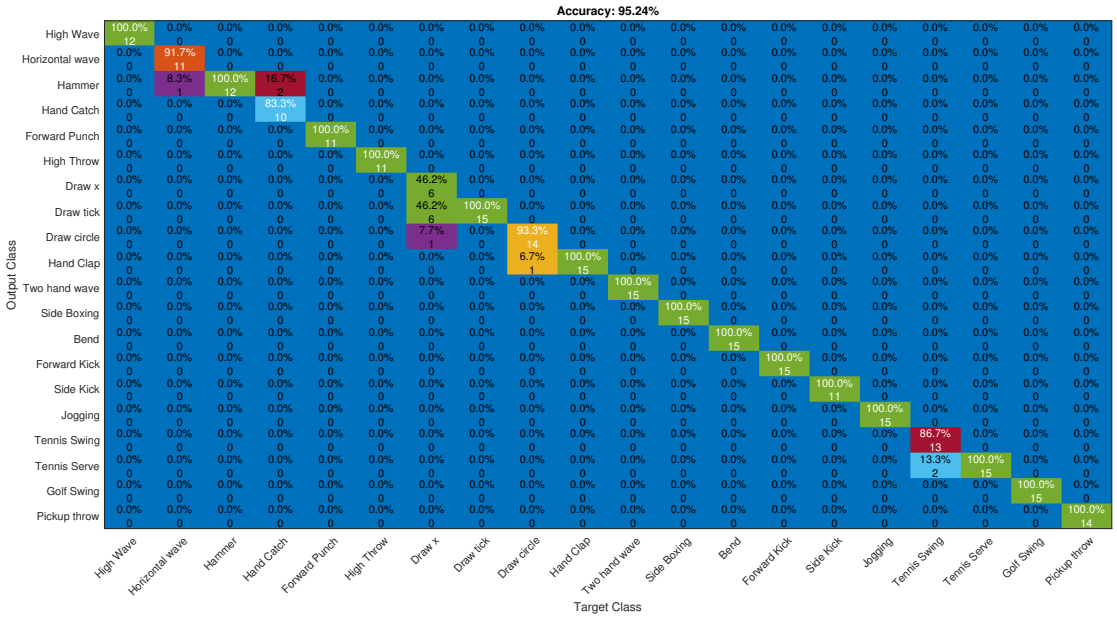


Figure 12: Confusion Matrix demonstrating the class-wise recognition rates under setting 2 for MSR Action 3D Dataset

Stacked Filter Bank based descriptor is designed using a pool of strong filters to filter this compact SDMM to extract the prominent features. The classification of the action sequences is achieved using a CRC. The proposed method is evaluated through extensive experiments on MSR Action3D dataset under standard protocols. Our experiments exhibit better results, with promising class-wise recognition improvement compared to state-of-the-art methods. Moreover, the proposed method generates significantly compact feature vector, with less computation compared to the UDTCWT feature descriptor. This

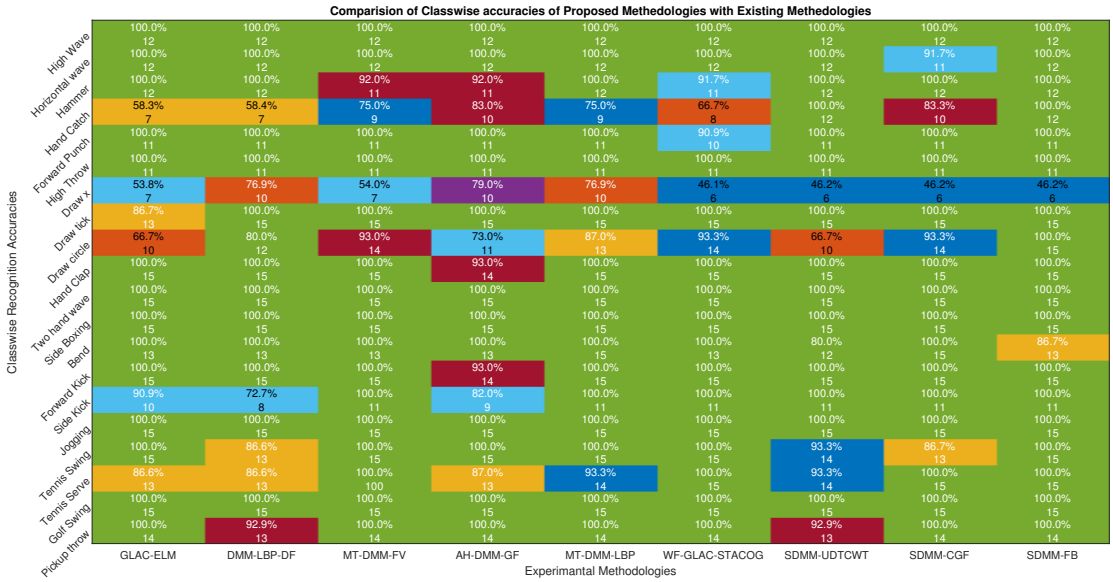


Figure 13: Comparison of recognition accuracies on Action3d dataset in Stacked Filter Based Model

work demonstrates the applicability of proposed feature descriptor for action recognition in a single subject experimental scenario, which can be extended to multi-subject scenarios as a future work. Through, Deep learning is an obvious choice for most of the computer vision applications, the demand of bulk-sized training data and high computational intensity, hinders the generic usage of deep networks. However, the proposed method can be trained with smaller dataset, with less computational complexity.

The proposed CGF based descriptor is very simple to implement and computationally inexpensive as no image convolutions are involved in the calculations of the complex gradient function. Further it can well capture the local variations as it is proved with the case studies of iris and depth sequence based human action recognition.

7.1 Contributions in our work

In the due course of our research work we introduce a compact and computationally efficient DMM representation, namely Stridden Depth Motion Map (S-DMM), which hastily generates an accumulated dense structure striding 4 frames of the video sequence at a time. This completes the DMM generation four times faster than the existing approaches. Also, we explore a Undecimated Dual Tree Complex Wavelet Transform (UDTCWT) based feature descriptor for extracting the wavelet features from our S-DMMs. We exploit the orientation and magnitude components of image gradient inorder to devise an efficient and simple feature descriptor. The image gradient is a robust property of gray scale images which exhibits important features of the image such as, illumination changes, the amount of change and the direction through which pixel value varies. In addition to that we develop a new Stacked Filter Bank(SFB) based descriptor which results in a significantly (roughly $\frac{1}{4}^{th}$ times) smaller and more discriminative feature vector. Motivated by the promising results achieved in some of the existing works, which exploit TSE and Riesz filters for other image processing applications, we propose a pool of filters based

partial differentials as the first and higher derivatives which are more efficient to intricate local features. Also, we explore Inception v3 network on RGBD video based HAR. We use a Inception v3 network pre-trained on ImageNet and adopt Transfer Learning to retrain Inception on our data and train our model on RGBD video data.

7.2 Publications:

1 **Action Recognition using Undecimated Dual Tree Complex Wavelet Transform from Depth Motion Maps/Depth Sequences**
Authors: B. H. Shekar, Rathnakara Shetty P., Sharmila Kumari M. and Leonid Mestetsky
Publication year: May 2019
Published in: ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W12, 2019. Conference: ISPRS International Workshop Photogrammetric and computer vision techniques for video Surveillance, Biometrics and Biomedicine, 13–15 May 2019, Moscow, Russia Pages: 203-209 Publisher: ISPRS
DOI: 10.5194/isprs-archives-XLII-2-W12-203-2019

Abstract: Accumulating the motion information from a video sequence is one of the highly challenging and significant phase in Human Action Recognition. To achieve this, several classical and compact representations are proposed by the research community with proven applicability. In this paper, we propose a compact Depth Motion Map based representation methodology with hastey striding, consisely accumulating the motion information. We extract Undecimated Dual Tree Complex Wavelet Transform features from the proposed DMM, to form an efficient feature descriptor. We designate a Sequential Extreme Learning Machine for classifying the human action secquences on benchmark datasets, MSR Action 3D dataset and DHA Dataset. We empirically prove the feasability of our method under standard protocols, achieving proven results

2 **Stacked Filter Bank based descriptor for Human Action Recognition from Depth Sequences**
Authors: Rathnakara Shetty P., B. H. Shekar, Leonid Mestetsky and Manju Prasad
Publication year: December 2019
Published in: Proceedings of the Conference on Information and Communication Technology Pages: 1-6 DOI: 10.1109/CICT48419.2019.9066216
Publisher: IEEE
Conference: Conference on Information and Communication Technology, IIIT Allahabad, India

Abstract: Registering the motion cues from a video to produce a compact representation is a crucial stage in video based Human Action Recognition (HAR). Exploiting the most prominent features using an efficient descriptor from such a representation also plays equally significant role in the performance of recognition models. In this work, we present a concise Depth Motion Map with striding which registers the motion cues from depth sequences on a video and a novel Filter Bank based descriptor,

wherein a Taylor Series Expansion (TSE) filter, a Riesz filter and a gradient filter are stacked together to extract the prominent features. We empirically evaluate the feasibility of our method on MSR Action 3D dataset under standard protocols, achieving state-of-the-art results.

3 **Complex Gradient Function Based Descriptor for Iris Biometrics and Action Recognition**
Authors: B H Shekar, Rathnakara Shetty P., and Sharada S. Bhat
Publication year: December 2020
Published in : Communications in Computer and Information Science book series (CCIS, volume 1377),
Pages: 489–501
DOI : 10.1007/978-981-16-1092-9-41
Publisher : Springer Nature Singapore Pte Ltd.
Conference: International Conference on Computer Vision and Image Processing (CVIP 2020), 4-6 December 2020, IIIT Allahabad, India

Abstract: Image gradient has always been a robust characteristic of digital image which possesses localised spatial information of each pixel in all the directions. Exploiting the gradient information at the pixel level is a good old technique applied in various fields of digital image processing. In this paper, the magnitude and direction of image gradient is explored to design a local descriptor. We propose a novel local feature descriptor based on Complex Gradient Function (CGF), which maps each pixel from the spatial plane into its complex extension involving the magnitude and direction of image gradient at that pixel. We exploit the proposed descriptor for human action recognition from depth sequences and human authentication using iris biometrics. The efficiency of descriptor is demonstrated with experimental results on benchmark datasets IITDelhi, MMU-v2, CASIA-Iris, UBIRIS, and MICHE-I for iris authentication and MSR Action 3D dataset for human action recognition(HAR).

The following works have been communicated to Journals and under review

1 Complex Gradient Function based Image Descriptor (Submitted to Springer Nature Computer Science SI: Recent Trends in Computer Vision Computer Science)

2 Wavelet Transform based descriptor for Human Action Recognition (Submitted to MDPI Journal of Imaging SI- Imaging Humans: Methods and Applications)

References

[1] Chen Chen, Zhenjie Hou, Baochang Zhang, Junjun Jiang, and Yun Yang. Gradient local auto-correlations and extreme learning machine for depth-based activity recognition. In *International Symposium on Visual Computing*, pages 613–623. Springer, 2015.

[2] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 1092–1099. IEEE, 2015.

[3] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of real-time image processing*, 12(1):155–163, 2016.

[4] Chen Chen, Mengyuan Liu, Hong Liu, Baochang Zhang, Jungong Han, and Nasser Kehtarnavaz. Multi-temporal depth motion maps-based local binary patterns for 3-d human action recognition. *IEEE Access*, 5:22590–22604, 2017.

[5] Chen Chen, Mengyuan Liu, Baochang Zhang, Jungong Han, Junjun Jiang, and Hong Liu. 3d action recognition using multi-temporal depth motion maps and fisher vector. In *IJCAI*, pages 3331–3337, 2016.

[6] Chen Chen, Baochang Zhang, Zhenjie Hou, Junjun Jiang, Mengyuan Liu, and Yun Yang. Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features. *Multimedia Tools and Applications*, 76(3):4651–4669, 2017.

[7] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[8] Z Gao, SH Li, YJ Zhu, C Wang, and H Zhang. Collaborative sparse representation leaning model for rgbd action recognition. *Journal of Visual Communication and Image Representation*, 2017.

[9] Zan Gao, Jian-ming Song, Hua Zhang, An-An Liu, Yan-bing Xue, and Guang-ping Xu. Human action recognition via multi-modality information. *Journal of Electrical Engineering and Technology*, 9(2):739–748, 2014.

[10] Tong Hao, Dan Wu, Qian Wang, and Jin-Sheng Sun. Multi-view representation learning for multi-view action recognition. *Journal of Visual Communication and Image Representation*, 48:453 – 460, 2017.

[11] J. F. Hu, W. S. Zheng, J. H. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[12] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015.

[13] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72(Supplement C):504 – 516, 2017.

[14] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. A depth video-based human detection and activity recognition using multi-features and embedded hidden markov models for health care monitoring systems. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4):54–62, 06/2017 2017.

[15] Michal Koperski and Francois Bremond. Modeling spatial layout of features for real world scenario rgb-d action recognition. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 44–50. IEEE, 2016.

[16] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[17] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE, 2010.

[18] Ivan Lillo, Juan Carlos Niebles, and Alvaro Soto. Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos. *Image and Vision Computing*, 59(Supplement C):63 – 75, 2017.

[19] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1053–1056. ACM, 2012.

[20] Hong Liu, Qinqin He, and Mengyuan Liu. Human action recognition using adaptive hierarchical depth motion maps and gabor filter. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 1432–1436. IEEE, 2017.

[21] Hong Liu, Lu Tian, Mengyuan Liu, and Hao Tang. Sdm-bsm: A fusing depth scheme for human action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4674–4678. IEEE, 2015.

[22] Zhi Liu, Chenyang Zhang, and Yingli Tian. 3d-based deep convolutional neural network for action recognition with depth sequences. *Image and Vision Computing*, 55:93–100, 2016.

[23] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 716–723, 2013.

[24] DS Rajesh and BH Shekar. Undecimated dual tree complex wavelet transform based face recognition. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, pages 720–726. IEEE, 2016.

[25] B. H. Shekar and S. S. Bhat. Steerable riesz wavelet based approach for iris recognition. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 431–436, Nov 2015.

[26] B H Shekar and Sharada S Bhat. Iris recognition using partial sum of second order taylor series expansion. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '16, pages 81:1–81:8, New York, NY, USA, 2016. ACM.

[27] B. H. Shekar, P. Rathnakara Shetty, M. Sharmila Kumari, and L. Mestetsky. Action recognition using undecimated dual tree complex wavelet transform from depth motion maps / depth sequences. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W12:203–209, 2019.

[28] Antonio W Vieira, Erickson R Nascimento, Gabriel L Oliveira, Zicheng Liu, and Mario FM Campos. On the improvement of human action recognition from depth map sequences using space–time occupancy patterns. *Pattern Recognition Letters*, 36:221–227, 2014.

[29] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *Computer vision–ECCV 2012*, pages 872–885. Springer, 2012.

[30] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297. IEEE, 2012.

[31] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012.

[32] X. Yang and Y. Tian. Super normal vector for human activity recognition with depth cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):1028–1039, May 2017.

[33] Xiaodong Yang and Ying Li Tian. Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 14–19, 2012.

[34] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060. ACM, 2012.

[35] Chong Zhao, Minglin Chen, Jinhao Zhao, Qicong Wang, and Yehu Shen. 3d behavior recognition based on multi-modal deep space-time learning. *Applied Sciences*, 9(4):716, 2019.

Research Scholar:

Rathnakara Shetty P.,
Research Scholar,
Department of Computer Science,
Mangalore University

Research Supervisor:

Dr. B.H. Shekar,
Professor,
Department of Computer Science,
Mangalore University